Explainable machine learning and time series analysis

January 17, 2022

1 Project description

The aim of this project is to develop a system that can help growers plan ahead for their crops by giving them information about the expected size of the crops in a future date. Using this system, they should be able to talk to supermarkets about when and how much produce they can sell them, resulting in a measurable gain for the growers' business. To do this, they have given you historical data from previous crops, as well as weather data.

2 Tasks

1. Read the dataset description below. Load, explore, clean, and preprocess the dataset. Some suggestions for preprocessing (but note this is not an exhaustive list and none of these is compulsory—except for the data leakage note—it's up to you!):
   - Edit column names: make sure that the column names can be comfortably used for handling data and operations such as merging data frames.
   - Add flight dates from *2020 planting data*.
   - Add missing planting dates from *2020 planting data* and other data that may be useful, e.g., *variety*, *comments*.
   - Figure out how to use date-time features. Most regressors need numerical variables, so you'll need to convert these to some other format. For example, you can create a variable called *days_to_check* that converts *flight_date* into number of days from *plant_date*.
   - Be very careful with **data leakage**: your task is to predict the size (see point 2 below) of the lettuce in the future, assuming today is *Flight time*: you cannot use information from the weather from the future!
   - Convert weather data into features. Hint: since you can't use data from the future, can you use an estimate using data from previous years?
   - Merge features you've created from the weather data with the plant dataframe so you can use them.
2. Using appropriate methodology, train and measure the performance of a multi-label linear regressor (you can play with the order of the polynomial) to simultaneously predict head weight, polar diameter, and radial diameter at the 'Check Date' using information about the plant at flight time and about the weather.
3. (Optional) Check performance also using other methods.
4. Make sure that you reflect on the interpretability of your methods and results (e.g., through feature importances).

3 Dataset description

The dataset attached contains data from crops and daily weather data from a weather station.

The tab named 'plants' contains measurement data, each row being a plant that was cut and measured. For each plant you have the following information:

- Batch number: the batch ID to which the plant belongs.
- Plant date: the date in which the batch that plant is part of was planted – You can ignore the rows for which no plant date is available.
- Flight date: the date in which the measurements were taken from the drone – You can fill the missing values for this column using the information from the "2020 Flight dates" tab.
- Check date: the date at which the manual measurements (weight and size) were taken.
- Measurement information:
  - Leaf Area (cm^2): leaf area measured from above by the drone on the day it was flown (typically 30-40% of the way into the growth cycle).
  - Leaves: the number of leaves in the plant.
  - Fresh Weight (g): the wet biomass of the plant. – Always available
  - Head Weight (g): the wet biomass of the plant in saleable form (i.e. as you'd find it pack in a box for the supermarket) – Do **not** use this column as a feature to make your predictions.
  - Polar Diameter (mm): the diameter of the lettuce head from base to top.
  - Radial Diameter (mm): the diameter of the lettuce head from side to side.
  - Diameter ratio: is computed from the polar and radial diameters. Do not use this column for your predictions.
  - Density (kg/L): is computed from some of the measures above. Do not use this column for your predictions.
  - **NOTE**: The head of the lettuce only forms about 50% through the growth cycle. Thus, for early measurements there is no information about the radial and polar diameters—instead, there is information about the fresh weight and the number of leaves. For later measurements we have fresh weight, head weight and polar and radial diameters.
- Class: category that the plant belongs to. They range from 1–5, with 1 being the smaller plants and 5 being the largest ones.
- Remove: drop the rows that are **NOT** blank in this column.

There are also two tabs that detail when each batch of crop was flown ("flight dates") as well as the planting records, which has the full details about each batch of crop ("planting"—you can use or ignore this sheet for your model).

The tab "weather" contains daily weather data from a local weather station.

Notes on measurements and dates

For each batch there is one measurement from the drone ("Leaf area", measured on "Flight date") and up to 3 manual measurements (depending on whether the head of the lettuce has been formed, either [fresh weight, head weight, polar diameter, radial diameter]) or [fresh weight, number of leaves])). The three manual measurements typically correspond to roughly 70%, 85%, and 98% of the growth cycle. For example, batch 517 was planted on 5th June, drone-measured on 8th July and manually measured on the 8th, 15th and 20th of July.

Note that the manual measurements require the plant to be cut, so they are done on different plants each time.